Review "Are flood damage models converging to reality? Lessons learnt from a blind test"

This paper compares many different flood damage models to a single validation case. This has been done before but since the last comparison a lot of new (supposedly better) models have become available that can be compared to each other. Furthermore, in previous studies and again in this study a lot of the results are surprising, and we clearly lack some understanding to explain all these results. I therefore believe this is a useful contribution to the literature because it provides yet more model comparison data and provides a lot of interesting speculation on the results. Especially the deep analysis of the limitations of the validation set are very interesting. However, the paper should be more cautious in the way it reasons in the results and discussion section and especially recognize the potential role of coincidences in the results.

- The paper is limited by there being only one validation set. This validation set has problems as is correctly mentioned in section 3.4. In earlier similar studies (e.g. Jongman et al., 2012) that used multiple validation studies it was common to see that some models performed well on one validation set and bad on another. I think this limitation of the study should be mentioned a bit more clearly, also to point out for future studies of this kind that it may be a good idea to have multiple validation sets (such as in Jongman et al., 2012). Because of this and the next point, the main value of the paper is in a comparison between the models rather than an absolute value judgement of the models.
- The second weakness is also not really highlighted and that is that the number of compared models is too small to draw any strong conclusions. Obviously, it wasn't feasible to select more models but some of the good model performances shown in this paper are pure coincidence, I'm sure about that. A good or bad performance of a model should therefore be seen as a single data point (sample) that could be just be a coincidence. This should be taken into account when drawing conclusions. I think this goes well in the conclusions section but in the results and in the discussion section some of the speculations should be done more cautious. Maybe go through these sections and reevaluate some of the reasoning based on the idea that there is a high level of coincidence in the results. Also be clearer about this limitation and the general idea that some observations could be simply a coincidence, add a few sentences about that somewhere.
- The statement in the discussion about the value of multi-variable models against simple models (line 526) cannot be stated like this. Multi-variable models can only be transferred when there is some overlap in the context between the training and validation data (see Wagenaar et al. 2008). When there is no overlap the transfer obviously wouldn't work no matter how many variables you add to the data. A multi-variable model should always be compared to a single-variable model based on the same data and not to a single-variable model from a different region to then conclude that a multi-variable model isn't useful. Maybe this could be done with a very large number of models but not based on the tiny number tested in this paper (especially because this observation also contradicts common sense and earlier findings elsewhere).
- The paper title, abstract and introduction puts a lot of emphasis on it being a blind-test. All properly carried out model validations studies don't use any knowledge from the validation data for the model development (standard practice). The constant emphasize on that in this study is therefore a bit misplaced I think. Did earlier studies not follow this approach? I

believe they did. Maybe they didn't advertise it this strongly, or maybe this study was more strict or systematic on that but does that really add anything special?

- This study is very similar to the paper Jongman et al., 2012. It might be interesting to add a paragraph in the discussion section to compare the results of the papers. Of course, this paper is comparing much newer and different models. Yet the general observation that the results are very different from each other and that its difficult to make sense of that is the same among the studies.
- The large group of authors suggests that an expert on each of the compared models was included in the paper. This is however clearly not the case for the Jonkman et al. model and hence the paper makes claims about this model that aren't true. Apart from fixing these mistakes (which I pointed out below), I think it should be clarified somewhere which experts worked on which models and whether the expert personally developed the model or interpreted information from literature (much more error prone).

Minor comments

Line 46: This is grey literature and I can't find it on the internet. Perhaps add one of the very many peer-reviewed journal articles that could also be used to support this statement and are often much older than 2007.

Line 101: I have never seen the term "low-variable" model, maybe consider a different word? Do you mean "single-variable" here?

In section 2.3/table 2. Some basic but important information in this section seems missing for some of the models, such as the origin of the model (country/region) (for Carisi et al. – mono) and the intended flood type (for many of them) and the year the model was made (maybe can be retraced from reference list but not easy for reader). Maybe its good to add this information to table 2 or at least make it clear in the section. Also maybe try to make the different texts a bit more uniform (i.e. present same information in same order).

Line 245. I don't see the significance of this model using a mathematical function. I also know much older models doing that and think its an irrelevant characteristic. Some other damage functions may also follow a mathematical function but just communicate it different.

Table 2: I don't think its correct to classify the Jonkman et al., 2008 model as empirical. It was inspired by many different sources of empirical data but no systematic empirical method was applied and the model is basically expert judgement considering a few empirical data points.

Section 3: I don't like this title very much why not just "results" this is confusing

Line 293-294: Maybe clarify that you look at a subset of the models here and didn't adjust the whole building models to only use the ground floor value (unclear at first).

Line 298 and 299: Can you rephrase this sentence its currently confusing.

Line 305: Could you rephrase this sentence its difficult to follow: "Individual damage estimates differ on average by a factor of 28, with the more frequent factor around 10."

Table 3: Could you split the 3th column, this is confusing and an uncommon form a presenting it.

The title of 3.1 is a bit unclear. Could you rephrase it? It's a bit long and I didn't get that with blind mode you mean that its not compared to observations yet (which I expected). Maybe call it "comparison of the models". Or keep the title and clarify in the first sentence that the reader shouldn't expect the comparison to real observations yet.

Figure 4: This table is interesting but requires some additional discussion. When you have two very different exposure values but the variation among the buildings solely depends on the size of the buildings the correlation between the two very different exposure values is still one. So this figure mostly says something about the characteristics that differ per building (if I understand the figure correctly). This doesn't become clear from the text.

Section 3.2. I like the content of this section but I had to read it twice to fully understand it. Maybe the authors could try to clarify this section a bit more. I think especially the title and the first sentence don't make it very clear at first (its all correct but it's a bit of deciphering for a reader).

Table 4: Could you split the column again like in the previous table.

Table 4: Could you consider including the model origin of all models in this table. That would be a very interesting reminder for the reader. Especially for readers who don't read the entire text this would be very useful.

Line 487: In case of the Netherlands this isn't true. The empirical data in this model is at best used relative. Absolute values are 100% synthetic.

Line 554: Micro-models are essential when measures are undertaken at micro level (e.g. for insurance or studies about elevating specific houses as is common in some countries). When aggregated damages are assessed they may indeed add less information. A second reason why micro-models are important is for at least for the location of buildings. The difference between a house flooding or not is sometimes a matter of just meters. Its therefore important to work with precise models on location even if the other building characteristics are all the same.